Lab 05 – clean and organize data to undertake statistical analysis

Learning Outcomes:

- Organize the temporary files in order to create the working file
- Complete the working file through the generation of new variables
- Undertake some simple descriptive statistics

Merging the temporary files

Two important commands to merge different temporary files:

`Merge` – to merge files with different series referred to the same observations (country / year):

> Achtung: all the temporary files that are to be merged have to be sorted in the same way. Moreover, we need to tell Stata what kind of merging is needed.

`Append` – to add observations to the file in use (the variables have to be the same, and with the same name).

Completing the working file

If the right variables have been chosen, if they have been checked and cleaned, if all the temporary files have been merged...

... we are not ready yet to undertake statistical analysis.

What else?

- Some new variables have to be generated from existing variables;
- Some existing variables have to be “aggregated”, since it might be necessary to compute the average value referred to a given period;
- Some existing variables have to be “aggregated” because they do not have yearly values.

Small exercise: the commands used are grouped in the file called: “Lab05.do”, that you find in the repository, together with files of example: Lab05a.txt Lab05b.txt Lab05c.txt.

Descriptive statistics

As a starting point, always prepare some basic tables of descriptive statistics

Descriptive statistics for all the variables included in the working file

Frequency tables for some of the variables included in the working file

Correlation tables for some of the variables included in the working file

Exercise – attempt to do the following operations on the test files Lab05a.txt, Lab05b.txt, Lab05c.txt, and save them in a do file

- Merge the temporary files in the final working file, and save it in .dta
- Compute the main descriptive statistics for every variable in the file, but only for the period including years from 3 to 5;
- Drop all the observations prior to year 2;
- Compute the average growth rate of consumption for the period including years 2-5;
- Estimate the model $CONS = a + b\ INC + \varepsilon$ on this subset of data.